Mots simples, mots complexes : étude lexicologique et création de ressources nouvelles

Núria Gala
LIF-CNRS, Aix Marseille Université, BLRI

Séminaire *Langues en Contact et Typologie*, Aix en Provence

03 avril 2014

Mots simples ? Mots complexes ?

"Les mots novlangues étaient divisés en trois classes distinctes (...). Le vocabulaire A comprenait les mots nécessaires à la vie de tous les jours, par exemple pour manger, boire, travailler, s'habiller, monter et descendre les escaliers, aller à bicyclette, jardiner, cuisiner, et ainsi de suite...
(...) Mais en comparaison avec le vocabulaire actuel, il y en avait un très petit nombre et leur sens était délimité avec beaucoup plus de rigidité. On les avait débarrasés de toute ambiguïté et de toute nuance."

G. Orwell *1984* (Paris, Gallimard, 1950, p.423)

Mots familiers... et fréquents

- Familiarité lexicale : [Dale, 1931] > constitution d'une liste de mots simples (dans une liste de 10.000 mots, ne sont retenus que ceux connus par au moins 80% des élèves, ce qui réduit la liste à 3.000 mots), Greenberg > liste des 100 mots présents dans toutes les langues
- Fréquence : [Grefenstette and Nioche, 2000] > listes de mots plus fréquents sur le Web (correlation avec longueur : plus fréquents > plus courts > (mais aussi plus ambigus !)

	English		Finnish		French		German
0.010	that	0.008	että	0.009	pour	0.030	die
0.010	is	0.008	ei	0.009	dans	0.029	und
0.008	it	0.006	oli	0.008	une	0.011	zu
0.007	you	0.005	tai	0.007	qui	0.010	von
0.007	with	0.005	joka	0.007	par	0.010	mit
0.007	be	0.005	myös	0.006	sur	0.009	für
0.006	this	0.004	mutta	0.006	est	0.009	ist
0.006	are	0.004	ovat	0.006	au	0.007	werden
0.005	or	0.004	kun	0.005	plus	0.007	auf
0.005	have	0.003	sen	0.004	pas	0.007	eine
0.005	not	0.003	ole	0.004	sont	0.007	im
0.005	we	0.003	niin	0.004	aux	0.006	nicht
0.004	will	0.003	kuin	0.004	avec	0.006	ein
0.004	they	0.003	hän	0.003	cette	0.006	sich
0.004	all	0.003	voi	0.003	vous	0.006	auch
0.003	his	0.003	sekä	0.003	nous	0.006	oder
0.003	but	0.003	jos	0.002	ont	0.005	sind
0.003	your	0.002	vain	0.002	leur	0.005	bei
0.003	their	0.002	sitä	0.002	ses	0.004	als
0.003	my	0.002	olla	0.002	ces	0.004	wird

Estimation of English and non-English Language Use on the WWW (2000)

Supercalifragilisticexpialidocius...

It's supercalifragilisticexpialidocious
Even though the sound of it is something quite atrocious
If you say it loud enough, you'll always sound precocious
Supercalifragilisticexpialidocious !

Richard & Ribert M. Sherman (1964)

Mots longs... et rares

Selon Wikipédia (http://en.wikipedia.org/wiki/Longest_words)

- Mots techniques (absents des dictionnaires de langue courante) :
  - pentakismyriahexakisquilioletracosiohexacontapentágono (ES, 54 lettres)
  - pneumonoultramicroscopicsilicovolcanoconiosis (EN, 45 lettres)
  - hexakosioihexekontahexaphobie (FR, 29 lettres)
- Mots courants :
  - anticonstitucionalmente / anticonstitutionnellement
  - otorrinolaringologia / oto-rhino-laryngologie > procédés de reccourcissement (troncation : "*tengo hora con el otorrino*" / siglaison : "j'ai rdv avec l'orl")

Premières idées...

Les critères de la longueur et de la fréquence correlés sont des indicateurs usuels pour estimer la complexité lexicale ; mais d'autres variables doivent être prises en compte :

- d'ordre psycholinguistique (familiarité, âge d'acquisition, etc.)
- d'ordre linguistique (structure des syllabes, taille de la famille morphologique, polysémie, etc.)

Dans cet exposé :

- Qu'est-ce qu'un mot simple/difficile ?
- La notion de complexité est-elle la même pour tous les publics ?
- Existent-ils des ressources pour ces mots-là ?
- Dans quelles applications ces ressources peuvent-ils être utiles ?

Sommaire

1. Sur la notion de complexité
   - Complexité linguistique
   - Complexité lexicale
2. Ressources lexicales
   - Dictionnaires informatisés et électroniques
   - Vocabulaires fondamentaux
   - Lexiques gradués
3. Lisibilité et simplification pour l'aide à la lecture
   - Domaine général
   - Applications concrètes
4. Conclusions

1 Sur la notion de complexité
- Complexité linguistique
- Complexité lexicale

2 Ressources lexicales

3 Lisibilité et simplification pour l'aide à la lecture

4 Conclusions

Complexité linguistique

Objectifs et problématique

- **Objectifs** : d'un point de vue typologique, comparer les langues en termes de complexité
  - hypothèse de l'equi-complexité (20e s.) : 'équilibrage' entre domaines au sein d'une langue
  - différents facteurs en diachronie et en synchronie contredisent cette hypohtèse : certaines langues atteignent les mêmes besoins de comm. avec moins de moyens (langues 'moins complexes', i.e. créoles)
- **Problématique** : identifier quels paramètres rendent une langue difficile
  - En typologie, comparer ce qui est comparable
  - En psycholinguistique, identifier qu'est-ce qui rend une production linguistique difficile à comprendre pour un public donné
  - En TAL, identifier qu'est-ce qui rend une production linguistique difficile à traiter computationnellement

Complexité linguistique

Approches dans la littérature

- Différentes approches, proposition de trois niveaux [Blache, 2011]
- Complexité globale : complexité d'une langue en tant que système, ex. longueur des productions, économie (nb catégories), irrégularités, sous-spécification (information provenant du contexte), etc.
- **Complexité locale** : complexité structurelle d'une réalisation, focus sur un sous-domaine, ex. profondeur d'un arbre syntaxique, nb de mots, etc.
- **Difficulté** : subjective, relative à un public donné, aspects liés au traitement, ex. dépendances incomplètes, adjectivation, etc.

Complexité lexicale

- Aspects linguistiques > apprentissage du vocabulaire (didactique)
- Domaine : lisibilité, prédiction objective de la difficulté lexicale
- Facteurs statistiques (fréquences)
- Facteurs intra-lexicaux (consistance phonème-graphème, régularités flexionnelles/dérivationnelles, nombre de sens par mot [Laufer, 1997])

Complexité lexicale

- Aspects psycholinguistiques > tâches de reconnaissance (décision lexicale, catégorisation)
- Idée principale : forte correlation entre fréquence et difficulté [Brysbaert et al., 2000]
- Autres facteurs : familiarité, âge d'acquisition, voisins orthographiques, nombre de morphèmes [Schreuder and Baayen, 1997] (reconnaissance visuelle de mots)

Déterminer la complexité lexicale (1/4)

Objectifs

- Déterminer des informations susceptibles de nous informer sur la complexité lexicale, les utiliser plus tard pour prédir automatiquement le niveau de difficulté d'un mot
- Calculer des valeurs statistiques sur des données
- Identifier des variables intra-lexicales extraites de ressources de différente nature

Ressource	**Type**	**Nb lemmes AANV**
Corpus patients parkinsoniens	parole pathologique	1.106
Corpus CEFR > FLELex	FR L2 (FLE)	17.870
Manulex	FR L1	19.037
Lexique 3	FR langue générale	47.342

Parole pathologique : corpus Pk

- Parkinson : troubles moteurs mais aussi dysarthrie (hypophonie, parole monotone, difficultés d'articulation) [Pinto et al., 2010]
- 20 enregistrements de patients 'off', 2.271 occurrences, 1.106 lemmes (AANV)
- Longueurs moyennes :
  - Corpus Pk > 6,3 lettres, 4,7 phonèmes, 1,96 syllabes
  - Lexique 3 > 8,6, lettres, 6,8 phonèmes, 2,9 syllabes
- Classement des mots du corpus selon Manulex :

	Niveau 1	**Niveau 2**	**Niveau 3**
Total corpus Pk	94,3%	1,45%	1,63%

Parole pédagogique : Manulex et corpus CERF

- Lexique avec vocabulaire classé en trois niveaux selon âge des apprentissages (CP, CE1, CE2-CM2)
- Corpus CERF, textes classés en six niveaux d'acquisition (A1, A2, B1, B2, C1, C2)

Niveau	Nb textes	Nb mots
A1	460	103.610
A2	487	166.680
B1	688	249.984
B2	203	130.752
C1	184	92.327
C2	49	34.482
Total	2.071	777.835

Nombre de textes et mots par niveau CERF

» Lexiques gradués, plus loin dans l'exposé...

Déterminer la complexité lexicale (2/4)

Ressources : Lexique 3, Manulex, corpus Pk

- Nombre de lettres, phonèmes, syllabes
- Structure syllabique (structures plus fréquentes dans corpus Pk : V, CVC, CV, CYV)
- Consistence graphème-phonème :
  - 0 = transparence : ’abruti’ `[abRyti]`
  - $<$ 2 caractères : ’abrite*r*’ `[abRite]`
  - $>$ 2 caractères : ’l*ent*em*ent*’ `[l@tm@]`
- Patrons orthographiques : doubles voyelles (ex. ée `[e]`), doubles consonnes (ex. pp `[p]`), digraphes (ex. ch `[S]`)

Déterminer la complexité lexicale (3/4)

Ressources : Morphalou, Manulex, Polymots, corpus Pk, corpus CEFR

- Morphèmes :
  - analyse morphologique automatique non supervisée, découpage en segments morphémiques étiquetés (base, préfixe, suffixe, élt. liaison)
  - probabilités transitionnelles entre des sous-chaînes observées dans le lexique [Bernhard, 2010]
  - nb morphèmes, préfixation (oui/non), suffixation (oui/non), est composé (oui/non), fréq. minimale préf/suf, fréq. moyenne préf/suf, taille famille morphologique

rouille – antirouille ; rouilleux
dérouiller – dérouillage ; dérouillement ;
débrouille – brouilleur ; brouilleuse ; débrouilleur ; débrouilleuse
brouille – brouillerie ; brouilleux

Déterminer la complexité lexicale (4/4)

Ressources : JeuxDeMots, BabelNet

- Polysémie :
  - utilisation de lexiques sémantiques (réseaux lexicaux)
  - a plusieurs sens dans *JeuxDeMots* (oui/non) (`http://www.jeuxdemots.org`) [Lafourcade, 2007]
  - nombre de synsets (groupes de synonymes) dans *BabelNet* (`http://babelnet.org/`) [Navigli and Ponzetto, 2010]

rouille(r_infopot#36 :25–> _INFO-POLYSEMIC) ['altération', 'rubigineux', 'sauce', 'érosion']
rouille(3) ///bn :00068634n|noun|rouille///bn :00068636n|noun|rouille///bn :00068637n|noun|champignon

Bilan

- Identification de 49 variables intra-lexicales [Gala et al., 2014] (*soumis*)
- Différents tests statistiques pour estimer quels sont les prédicteurs qui apportent plus d'informations sur la difficulté lexicale :
  1.
  2.
  3.
  4.
  5.
  6.
  7.
  8.
  9.
  10.

Bilan

- Identification de 49 variables intra-lexicales [Gala et al., 2014] (*soumis*)
- Différents tests statistiques pour estimer quels sont les prédicteurs qui apportent plus d'informations sur la difficulté lexicale :
  1. nombre de phonèmes
  2. polysémie
  3. nombre de lettres
  4. nombre de syllabes
  5. nombre de voisins orthographiques
  6. présence de voyelles nasales
  7. taille famille morphologique
  8. préfixation
  9. nombre de morphèmes
  10. patrons orthographiques (doubles voyelles, doubles consonnes, digraphes)

1 Sur la notion de complexité

2 Ressources lexicales
- Dictionnaires informatisés et électroniques
- Vocabulaires fondamentaux
- Lexiques gradués

3 Lisibilité et simplification pour l'aide à la lecture

4 Conclusions

Ressources lexicales

Définitions

1. Recueil de mots classés en liste pour différents usages.
2. Répertoire d'unités lexicales mises en parallèle pour deux langues.
3. Ensembles structurés de mots, quel que soit leur support, avec des informations associées à ces mots.

- Approche sémasiologique (des mots aux sens) / approche onomasiologique (des sens aux mots).
- Ressources lexico-sémantiques : informations linguistiques et sémantiques.

Trois évènements historiques marquants

1. Naissance de l'écriture (-3300 a.v. J.-C.) : tablettes d'Ebla (-2600 a.v. J.-C.), Liber Glossarum (8e siècle)
2. Invention de l'imprimerie (15e siècle) : Nebrija (1492), Estienne (1539), Percyvale (1591)
3. Développement de l'informatique (20e siècle) : TLFi (1957, 2000), WordNet (1990), BabelNet (2010)

Évolutions des besoins humains et évolutions techniques > implications dans la *conception* et la *diffusion* des ressources lexicales [Gala, 2013].

Dictionnaires informatisés (1/2)

Du papier à l'écran : les mots cliquables [Gala, 2013]

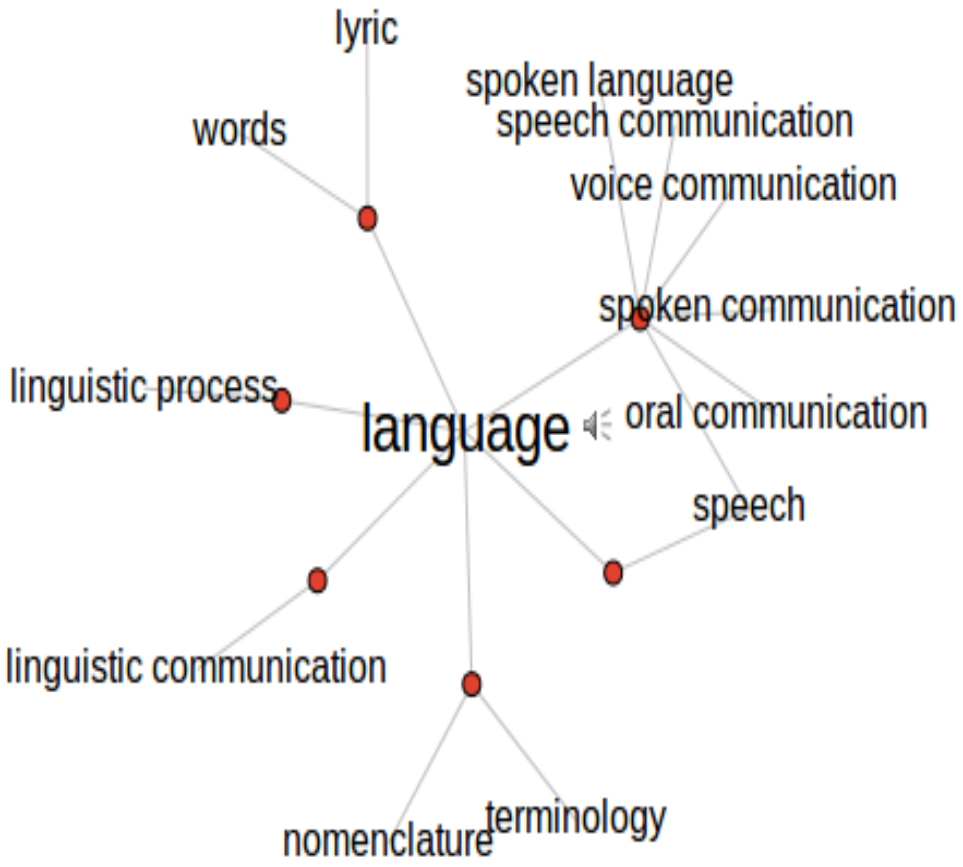
- Informatisation des données et des méthodes lexicographiques.
- Traitements de corpus pour la constitution des dictionnaires.
- Formats structurés : balisage SGML puis XML.
- Données multimédia dans des dictionnaires spécialisés :
  - phonétique (prononciation : `http://www.howjsay.com/`)
  - banques d'images (dessins, photos, icônes : `http://www.catedu.es/arasaac`)
  - images animées (ressources pour les langues de signes : `http://www.sematos.eu`)
- Exemples : *TLF/Frantext* (1957-1994), *Longman's* (1978), *Collins COBUILD* (1987), versions CD-ROM et en ligne du *Robert* (1996, 2001), *Larousse* (2010), etc.

Dictionnaires informatisés (2/2)

- Informatisation des moyens d'accès aux dictionnaires et à leurs contenus.
- Nouvelles possibilités de consultation des entrées (mots clés, propriétés morphologiques et/ou syntagmatiques, navigation, etc.).
- Visualisation statistique (nuages de mots) et diagrammatique (réseaux)
- Nouvelles plateformes, dictionnaires de dictionnaires et autres ressources (thesaurus, corpus, listes selon critères divers, etc.) :
  - *Nuevo Tesoro Lexicográfico de la Lengua Española* (`http://buscon.rae.es/ntlle`)
  - *Wordnik* (`http://www.wordnik.com/`)

Exemple (1/2)

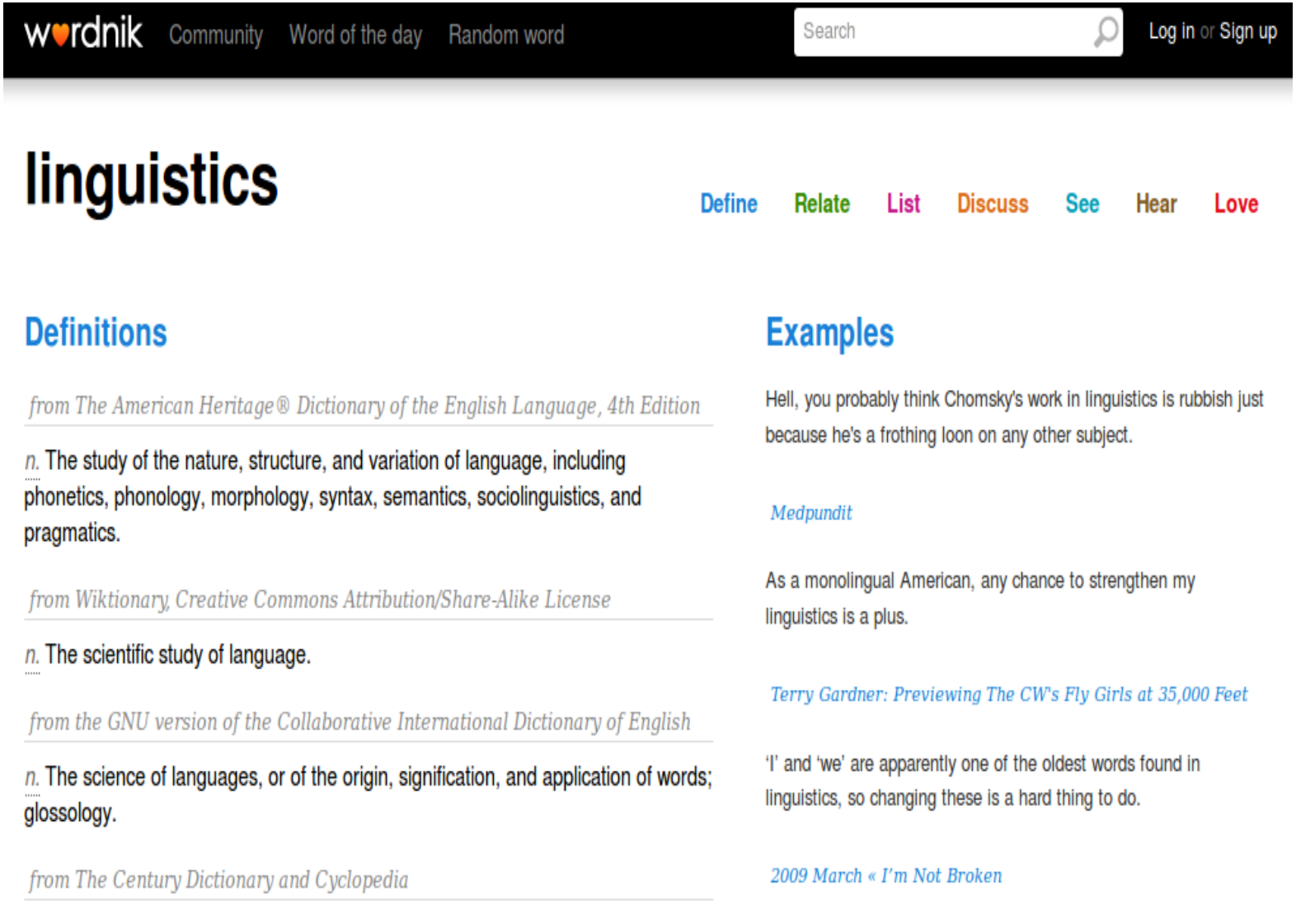
Visualisation en réseau (*net-like*)



Visual Thesaurus

Exemple (2/2)

Plateforme *all-in-one*



Wordnik

Ressources électroniques (1/2)

Pour l'homme...

- Ressources sans version papier au préalable, ex. le *TLFi*, le *Wiktionnaire*...
- Possibilités nouvelles au niveau des ressources multilingues (vrai multilinguisme possible dans une seule ressource, ex. *Reverso, Freelang*...) et multimodales (pour les langues de signes, ex. *Sematos*)
- Plateformes multi-ressource pour le traducteur, pour l'apprenant (ex. *Alfalex* [Selva et al., 2004]), pour le rédacteur professionnel (ex. *Antidote*), etc. et multi-support

Ressources électroniques (2/2)

Pour l'homme... et surtout pour la machine

- Ressources pour le traitement automatique des langues (TAL) : information explicite et structurée exploitable par l'ordinateur
- Contenu diversifié :
  - morphophonologie (*Celex* [Baayen et al., 1995])
  - syntaxe (*DicoValence* [Van den Eynde and Mertens, 2006]
  - sémantique (*WordNet* [Miller, 1990], *FrameNet* [Baker et al., 1998])
  - vocabulaire (*Manulex* [Lété et al., 2004])
- Destinées à différentes applications : analyseurs morpho-syntaxiques, extracteurs d'information, systèmes de question réponse, traducteurs et résumeurs automatiques, etc.

Quelques constats

- Hétérogénéité de ressources et de contenus
- Intégration dans des plateformes assistives
- Informatique > améliorations au niveau des accès aux données lexicales et à leurs informations

Mais...

- Disponibilité et utilisation en TAL de ces ressources pas toujours évidente
- Information sur la complexité lexicale inexistante

Méthodologie de construction

J. Murray. The Oxford English Dictionary. First edition. (fin 19e s.).

- Construction manuelle (première moitié 20e s.)
- Utilisation de grands corpus, statistique lexicale (deuxième moitié 20e s.)
- Plus récemment > techniques de TAL, combinaison de critères : corpus, analyse linguistique, analyse statistique, apprentissage automatique, contributions (*crowdsourcing*, "myriadisation du travail parcellisé")

Vocabulaires fondamentaux

Définition

Liste du vocabulaire minimal d'une langue.

- Nouvelles approches pour l'enseignement du lexique (début 20e siècle)
- Idée controversée, simplification du lexique idée réductrice
- Application à des formules de lisibilité
- Approches 'logiques', procédés rationnels : *Basic English* [Ogden, 1930]
- Méthodes statistiques : *Teacher's Book of Words* [Thorndike, 1921] et [Thorndike, 1944]
- Combinaison d'approches, fréquences obtenues pour des mots 'disponibles' (familiers) : *Français Fondamental* [Gougenheim, 1958]

Lexiques gradués

Ressources nouvelles

Définition

Ressources lexicales où les mots ont un niveau de difficulté associé, calculé en fonction de différents paramètres.

- *Manulex* [Lété et al., 2004]
- *ReSyf* [Gala et al., 2013]
- *FLELex* [François et al., 2014]

Manulex

[Lété et al., 2004]

- Liste de mots créée à partir de 54 manuels scolaires (français L1), 19.037 lemmes lexicaux (et 4.863 grammaticaux)
- Classement en trois niveaux selon leur apparition dans les manuels de la première année de primaire, de la deuxième ou des trois suivantes (selon volume d'acquisition de vocabulaire)
- Différentes mesures statistiques (fréquence, dispersion, etc.)

Mot	Cat.	Niveau 1	Niveau 2	Niveau 3
pomme	N	724	306	224
vieillard	N	-	13	68
patriarche	N	-	-	1
cambrioleur	N	2	-	33
Total dans Manulex		31%	21%	48%

ReSyf

[Gala et al., 2013]

- REssource lexicale en français avec des SYnonymes gradués selon leur niveau de diFficulté
- Construite semi-automatiquement à partir de différentes ressources existantes :
  - Lexicaux : *Lexique 3* [New et al., 2001], *Manulex* [Lété et al., 2004], *JeuxDeMots* [Lafourcade, 2007]
  - Textuels : corpus Parkinson (étude de la parole pathologique)
- Liste initiale : 19.037 lemmes lexicaux de Manulex (AANV), transformation en trois classes selon attestation dans un niveau scolaire de Manulex
- Liste finale (version 2013) : 12.687 lemmes lexicaux de Manulex avec des synonymes dans JdM

Exemple : données du lexique ReSyf

renard(n1) : canidé(n3) malin(n1) futé(n1) goupil(n2) roublard(n3)
pourtant(n1) : cependant(n1) néanmoins(n2) seulement(n1) toutefois(n2)
armure(n1) : cuirasse(n2) tissage(n3) harnais(n3) protection(n1)
piétiner(n2) : fouler(n3) piaffer(n3) trépigner(n1) marcher(n1)
glacial(n2) : impassible (n3) imperturbable(n3) rigoureux(n2) inhospitalier(n3) sec(n1) froid(n1) insensible(n3) glacé(n1) polaire(n2)
patriarche(n3) : chef(n1) vieillard(n2) père(n1)
joncher(n3) : couvrir(n1) parsemer(n2) tapisser(n1) disséminer(n3) recouvrir(n1)
policier(n1) : poulet(n1), flic(n2), commissaire(n3)
extravagance(n3) : absurdité(n3) folie(n1) bizarrerie (n3) frasque(n2) caprice(n1) excentricité(n3) originalité(n3) démence(n3) fantaisie(n2)

» Travail en cours et à venir : calcul du niveau de difficulté pour les mots absents de Manulex, désambiguïsation sémantique, niveaux de langue.

FLELex

[François et al., 2014]

- Liste de mots créée à partir de corpus CEFR de 777.835 mots (FR L2)
- Extraction de 16.833 lemmes lexicaux (et 1.038 grammaticaux)
- Segmentation et étiquetage morphologique automatique avec des outils de TAL (*tokenization, tagging*)
- 31% entrées avec fréq > 10 (dont 6% avec fréq > 100) et 69% entrées avec fréq < 10 (dont 20% d'hapax)
- Comparaison à Lexique 3 (47.342 lemmes langue générale) : 622 entrées de FLELex absentes de Lexique3 (3,5%)

Exemple : données du lexique FLELex

lemme	A1	A2	B1	B2	C1	C2
voiture	633.3	598.5	482.7	202.7	271.9	25.9
abandonner	35.5	62.3	104.8	79.8	73.6	28.5
justice	3.9	17.3	79.1	13.2	106.3	72.9
kilo	40.3	29.9	10.2	0.0	1.6	0.0
piétiner	0.0	0.39	0.0	0.53	15.7	0.0
logique	0.0	0.0	6.8	18.6	36.3	9.6
absurdité	0.0	0.0	0.34	4.55	3.29	67.36
en bas	34.9	28.5	13	32.8	1.6	0.0
en clair	0.0	0.0	0.0	0.0	8.2	19.5
de surcroît	0.0	0.0	0.0	0.0	15.67	0.0
donner rendez-vous	0.53	0.69	1.89	0.0	0.0	0.0
donner naissance	0.0	0.25	0.0	0.0	0.0	4.12

1 Sur la notion de complexité

2 Ressources lexicales

3 Lisibilité et simplification pour l'aide à la lecture
- Domaine général
- Applications concrètes

4 Conclusions

Lisibilité et simplification

- Lisibilité computationnelle : prédiction automatique objective de la difficulté d'un texte (formules de lisibilité)
- Simplification : processus de transformation d'un texte en un équivalent plus compréhensible étant donné un public donné
- Travaux en TAL (principalement en lisibilité textuelle) :
  - algorithmes statistiques (lisibilité = problème de classification : classer les documents (ou les mots) dans des niveaux de lisibilités donnés)
  - exploitation de variables lexico-syntaxiques de surface [Collins-Thompson and Callan, 2005] ou des arbres syntaxiques [Zhu et al., 2010]

Public(s)

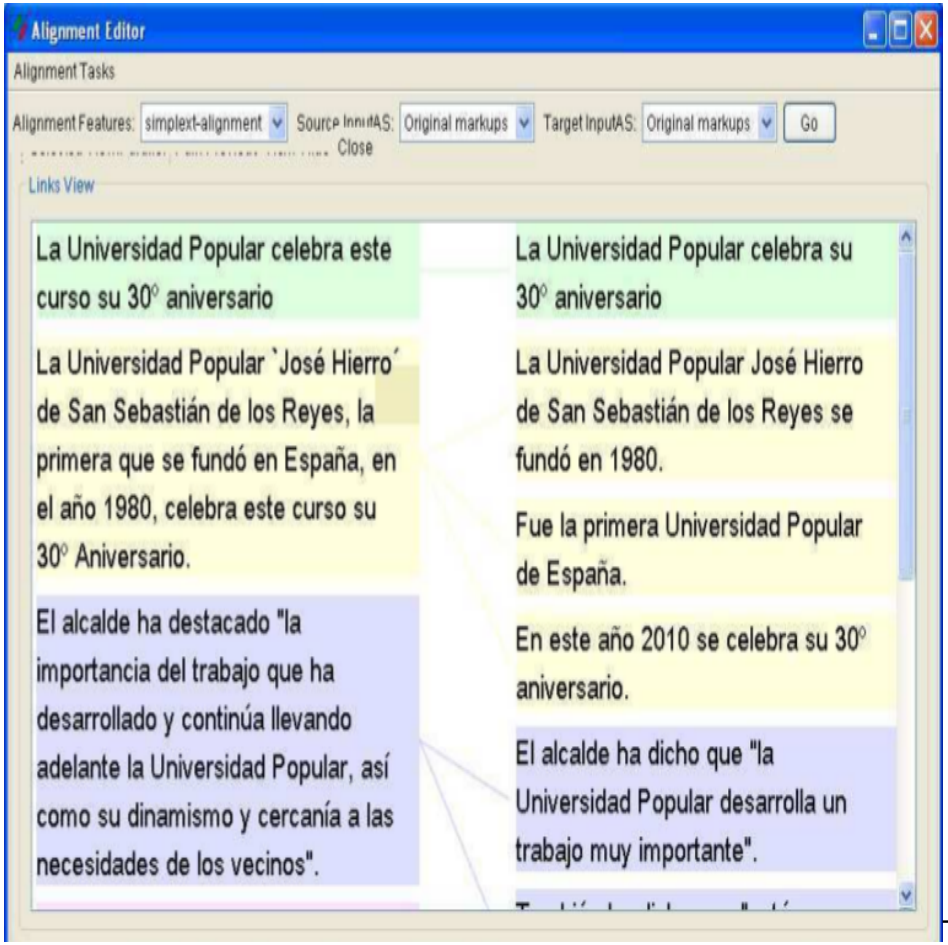
Complexité > Difficulté [Blache, 2011]

- Identification de classes d'individus (profils de lecteurs)
  - apprenants L1 en milieu scolaire [Lété et al., 2004]
  - apprenants L2 (FLE) [François, 2011]
  - personnes avec peu d'instruction [Watanabe et al., 2009]
  - parole pathologique : dyslexie [Rello et al., 2013], aphasie [Carroll et al., 1998], écrit des sourds...
- Besoins sociétaux, applications à la parole pédagogique et pathologique
- Domaine en activité croissante en TAL (workshop NAACL HLT 2012, H. Saggion RANLP 2013, *special issue of the International Journal of Applied Linguistics 2014*, *3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations EACL 2014*)

Aide à la lecture

Problématiques de recherche

- Dévelopement d'outils et de ressources pour l'aide à la lecture (*peut-on transformer automatiquement un texte en son équivalent plus simple ? quelle méthode ?*)
- Adaptativité à des publics spécifiques (*est-il possible d'utiliser un système mis en place pour un type de public pour un autre public ? d'une langue pour une autre langue ?*)
  - Handicaps langagiers et/ou cognitifs (*quelles particularités ?*) :
    - Autisme (projet FIRST > anglais)
    - Dyslexie (projet SIMPLEXT > espagnol)
    - Illetrisme (projet PorSimples > portugais)
  - Apprenants L1 ou L2



Text Simplification in Simplext : Making Texts more Accessible, Saggion et al. (2011)

- Projet SILK : SImplication pour l'aide à la Lecture et la Compréhension (public visé : enfants en difficulté, langue : français) » ... à venir

1 Sur la notion de complexité

2 Ressources lexicales

3 Lisibilité et simplification pour l'aide à la lecture

4 Conclusions

Conclusions

- Complexité lexicale > vers un modèle capable de prédire le niveau de difficulté en fonction de paramètres intra-lexicaux
- Ressources lexicales électroniques > vers des lexiques gradués intégrant la notion de difficulté
- Besoins sociétaux pour des publics spécifiques > parole pédagogique et pathologique
- Applications de TAL pour l'aide à la lecture

"**La langue** étant ce qu'elle est, de quelque côté qu'on l'aborde, on n'y trouvera **rien de simple** ; partout et toujours ce même équilibre complexe de termes qui se conditionnent réciproquement."

F. de Saussure. *Cours de linguistique générale*. pp.168-169.

Merci de votre attention !

Baayen, R. H., Piepenbrock, R., and van Rijn, H. (1995).
The Celex lexical database (Release 1) [CD-ROM].

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998).
The Berkeley FrameNet project.
In *COLING-ACL 98 : Proceedings of the Conference*, pages 86–90, Montreal, Canada.

Bernhard, D. (2010).
Apprentissage non supervisé de familles morphologiques : comparaison de méthodes et aspects multilingues.
Traitement Automatique des Langues, 2(51) :pp. 11–39.

Blache, P. (2011).
A computational model for language complexity.
In *1st Conference on Linguistics, Biology and Computational Science*, Tarragona, Spain.

Brysbaert, M., Lange, M., and Wijnendaele, I. V. (2000).
The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition : Further evidence from the Dutch language.

European Journal of Cognitive Psychology, 12(1) :65–85.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998).
Practical simplification of English newspaper text to assist aphasic readers.
In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*.

Collins-Thompson, K. and Callan, J. P. (2005).
Predicting reading difficulty with statistical language models.
JASIST, 56(13) :1448–1462.

Dale, E. (1931).
A comparison of two word lists.
Educational Research Bulletin, 18(10) :484–489.

François, T. (2011).
Les apports du traitement automatique du langage à la lisibilité du français langue étrangère.

François, T., Gala, N., Watrin, P., and Fairon, C. (2014).
FLELex : a graded lexical resource for French foreign learners.

In *Proceedings of International conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.

Gala, N. (2013).
Ressources lexicales mono- et multilingues : une évolution historique au fil des pratiques et des usages.
In *Ressources Lexicales. Contenu, construction, utilisation, évaluation.*, volume 30, pages 1–42. John Benjamins, Amsterdam, Gala, N. et Zock, M. edition.

Gala, N., François, T., Bernhard, D., and Fairon, C. (2014).
Un modèle pour prédire la complexité lexicale et graduer les mots.
In *Actes de TALN 2014 (soumis)*, Marseille, France.

Gala, N., François, T., and Fairon, C. (2013).
Towards a French lexicon with difficulty measures : NLP helping to bridge the gap between traditional dictionaries and specialized lexicons.
In *E-lexicography in the 21st century : thinking outside the paper.*, Tallin, Estonia.

Gougenheim, G. (1958).

Dictionnaire fondamental de la langue française.
Didier, Paris.

Grefenstette, G. and Nioche, J. (2000).
Estimation of English and non-English Language Use on the WWW.
CoRR, cs.CL/0006032.

Lafourcade, M. (2007).
Making people play for Lexical Acquisition.
In *Proc. SNLP 2007, 7th Symposium on Natural Language Processing.*,
Pattaya, Thailande.

Laufer, B. (1997).
What's in a word that makes it hard or easy : Some intralexical factors that affect the learning of words.
Cambridge University Press.

Lété, B., Sprenger-Charolles, L., and Colé, P. (2004).
Manulex : A grade-level lexical database from French elementary-school readers.
Behavior Research Methods, Instruments and Computers, 36 :156–166.

Miller, G. A. (1990).
WordNet : An On-Line Lexical Database.
International Journal of Lexicography, 3(4).

Navigli, R. and Ponzetto, S. P. (2010).
BabelNet : building a very large multilingual semantic network.
In *48th annual meeting of the Association for Computational Llinguistics.*, pages 216–225, Uppsala, Suède.

New, G. A., Pallier, C., Ferrand, L., and Matos, R. (2001).
Une base de données lexicales du français contemporain sur Internet : Lexique 3.
L'année psychologique, 101 :447–462.

Ogden, C. K. (1930).
Basic English : A General Introduction with Rules and Grammar.
Paul Treber, London.

Pinto, S., Ghio, A., Teston, B., and Viallet, F. (2010).
La dysarthrie au cours de la Maladie de Parkinson. Histoire naturelle de ses composantes : dysphonie, dysprosodie et dysarthrie.

Revue Neurologique, 166(10) :800–810.

Rello, L., Baeza-Yates, R., and Saggion, H. (2013).
The impact of lexical simplification by verbal paraphrases for people with and without Dyslexia.
Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science, 7817 :501–512.

Schreuder, R. and Baayen, H. (1997).
How complex simplex words can be.
Journal of Memory and Language, pages 118–139.

Selva, T., Verlinde, S., and Binon, J. (2004).
ALFALEX, un environnement d'aide à l'apprentissage lexical du français langue étrangère.
In *Congrès de l'ACFAS*, Montréal.

Thorndike, E. (1921).
The Teacher's Word Book.
Teachers College, Columbia University, New York.

Thorndike, E. . L. I. (1944).

The Teacher's Word Book of 30,000 words.
Teachers College, Columbia University, New York.

Van den Eynde, K. and Mertens, P. (2006).
Le dictionnaire de valence DicoValence, manuel d'utilisation.

Watanabe, W. M., Junior, A. C., Uzêda, V. R., Fortes, R. P. d. M., Pardo, T. A. S., and Alusio, S. M. (2009).
Facilita : reading assistance for low-literacy readers.
In *SIGDOC '09 : Proceedings of the 27th ACM international conference on Design of communication*, pages 29–36, New York, NY, US.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010).
A monolingual tree-based translation model for sentence simplification.
In *Proceedings of The 23rd International Conference on Computational Linguistics.*, Beijing, China.